

RESEARCH ARTICLE

Multi-Scalar Spectral Clustering: A New Approach to Capture Scale-Dependent Persistent Clusterings

FRANCIS BAFFOUR-AWUAH JUNIOR^{1,2}, MEHRAN FAZLI², AND DEBORAH A. STRIEGEL²¹Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA²Austere environments Consortium for Enhanced Sepsis Outcomes (ACESO), The Henry M. Jackson Foundation for the Advancement of Military Medicine Inc., Bethesda, MD 20817, USA

Corresponding author: Deborah A. Striegel (dstriegel@aceso-sepsis.org)

ABSTRACT Clustering is an unsupervised learning technique that is used to group a given set of data, typically requiring careful parameter tuning to achieve meaningful results. Spectral clustering, based on pairwise similarities between data points, is sensitive to the scale parameter σ in the commonly used Radial Basis Function (RBF) kernel. This paper introduces a new multiscale spectral clustering algorithm that identifies persistent, scale-dependent clustering structures by applying the eigengap heuristic across a continuous range of σ values. The method is validated using synthetic and real-world datasets. In addition, cluster quality is assessed using both external and internal performance metrics. The proposed algorithm detects robust clusterings at multiple scales, capturing global patterns and locally homogeneous groupings. This approach reveals significant structures that may be missed by single-scale methods.

INDEX TERMS Clustering, spectral clustering, eigengap, multiscale, persistent regions, RBF kernel.

I. INTRODUCTION

Clustering algorithms, such as spectral clustering, are widely used unsupervised learning techniques known for their ability to identify complex patterns in data. These methods serve as powerful exploratory tools and are applied across a wide range of fields, including image segmentation [1], bioinformatics, marketing, and social network analysis. For example, spectral clustering has been used to cluster biological sequence data [2], as well as to identify protein complexes in protein-protein interaction networks [3]. In social networks, it has been employed to assess behavioral patterns of students in an e-Learning environment [4]. In the medical field, clustering has been applied to identify distinct sepsis subgroups to inform more personalized treatment strategies [5].

There are many different approaches to clustering. Centroid-based algorithms, such as the K-means algorithm [6], assign data points to a predetermined number of clusters by minimizing the sum of squared distances between data

points and their respective cluster centroids. This algorithm, however, requires the number of clusters to be specified in advance, is sometimes not robust to outliers, and can be inefficient at capturing non-convex clusters. Density-based algorithms, on the other hand, such as the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [7], identify clusters as neighboring regions of high densities separated by sparse regions. DBSCAN does not require a specific number of clusters to be defined a priori; however, it can be ineffective for clustering datasets with large variations in density. Distribution-based algorithms, such as the Gaussian Mixture Model, identify clusters as sets of objects that are likely to have been generated from the same underlying probability distribution [8]. These algorithms may not yield optimal results for many real-world datasets where the underlying data distribution does not align with an assumed distribution, such as a Gaussian or binomial distribution.

Graph-based clustering algorithms transform data into graphs, with nodes or vertices representing data points and edges denoting connections (e.g., similarity as determined by a given similarity metric) between points. Representing

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro¹.

data with graphs allows the local and global properties of the data to be analyzed using graph-theoretical concepts, such as local neighborhoods and global connectivity [9]. Spectral clustering, a graph-based clustering algorithm, determines balanced graph partitions using a similarity metric to cluster observations [10]. Spectral clustering does not make any strong assumptions about the shape or size of clusters compared to the K-means and DBSCAN algorithms. Hence, it is suitable for capturing non-convex clusters (e.g., intertwined spirals, concentric circles) [11], as well as clusters of varying densities. Moreover, spectral clustering avoids the complex problem of identifying the underlying distribution of the data, as information about the clustering structure is encoded in the spectrum of the similarity graph.

In recent years, ensemble clustering approaches have become popular [12], [13], [14]. These methods evaluate the consistency of sample assignments across multiple clustering results. They distinguish between core samples, which are stably clustered, and halo samples, which are more uncertain. The algorithm first clusters only the core samples to form a strong consensus structure, then incrementally assigns the halo samples based on their similarity to the core. This strategy has been shown to improve clustering robustness and accuracy, particularly in complex or noisy datasets.

Most clustering algorithms require that the number of clusters be determined in advance. Two widely used methods for cluster-number determination are the Akaike Information Criterion (AIC) [15], and the Bayesian Information Criterion (BIC) [16]. An advantage of using these goodness-of-fit measures is that they are simple to compute. In their implementation, the AIC, the BIC, or both are computed and plotted as functions of the number of clusters. The “elbow” of the resulting curve is typically used to determine the optimal number of clusters [17], [18]. However, in many real-world applications, choosing the optimal elbow point can be difficult especially when the graph does not contain a sharp elbow.

Spectral clustering takes a data-driven approach called the eigengap heuristic (described below) proposed by [10], and based on perturbation theory to suggest the optimal number of clusters. This heuristic uses the difference between successive eigenvalues of a similarity matrix to determine the optimal number of clusters of a dataset. One commonly used similarity metric to construct the similarity matrix is the Radial Basis Function (RBF) kernel (Eq. 1). The scaling parameter σ determines the region of similarity surrounding each data point and can affect the estimated number of clusters detected within the data. Another complication arises when clustering data distributed across varying scales. One suggested solution involves iterating through a range of scale values and choosing the one associated with the tightest clusters [19]. However, this approach can be computationally expensive and may prove ineffective in datasets containing multiple dense and sparse regions of varying sizes, as there may not exist a single optimal scale that adequately captures

the structure across the entire dataset. In response to this, Zelnik-Manor and Perona [20], proposed a solution in which a specific scaling parameter is computed for each data point, depending on the local characteristics of its neighborhood, as defined by its K nearest neighbors. In this case, the scale for a point $\mathbf{s}_i$ can be defined as $\sigma_i = d(\mathbf{s}_i, \mathbf{s}_K)$, where $\mathbf{s}_K$ is the K -th neighbor of $\mathbf{s}_i$. Although this method allows for self-tuning of pairwise distances based on the statistical properties of each point’s local neighborhood, selecting an appropriate value of K to define that neighborhood can be challenging. Other methods for clustering exploit the spectral characteristics of block diagonal affinity matrices to automatically choose a scale for the dataset [21].

In practice, real-world datasets often exhibit structure at multiple levels of granularity. As such, a single clustering, regardless of how scale is defined, may fail to capture the full complexity of the data.

Multiscale approaches for spectral clustering have been developed. Some approaches involve using random walks to predict several plausible ways of clustering a dataset, and assigning scores to resulting groupings [22], [23]. Other approaches estimate the optimal number of clusters by computing the spectrum of the Laplacian of the data over an interval of scale values, and finding the largest eigengap spanning every index and scale [24]. Again, another approach involves recursively applying the eigengap heuristic to uncover clusters at multiple scales within a hierarchical tree structure, an algorithm known as Iterative Eigengap Search (IES) [25].

A number of multiscale spectral clustering methods for image segmentation have also used hierarchical structures, superpixels, and stochastic techniques to reduce computational complexity while preserving segmentation accuracy, enabling efficient processing of high-resolution and large-scale images [26], [27], [28].

In this study, a new method for determining optimal and robust scale-dependent clusterings, Multi-scalar Spectral Clustering (MSSC), is proposed. By observing how the number of clusters changes as the scaling parameter is varied, one can identify a prominent set of scale-dependent cluster counts within the dataset. This viewpoint offers multiple benefits. First, regions of the scale space where the number of clusters is unchanged across multiple scales, referred to as “persistent regions”, can be identified. The size of a persistent region provides a measure of the robustness of the number of clusters within that region for a given dataset. Second, within a persistent region, the stability of individual data points within clusters can be evaluated. This approach demonstrates that, in certain datasets, there may not be a single optimal clustering, but rather a set of robust, scale-dependent clusterings.

II. THEORY OF SPECTRAL CLUSTERING

The origins of spectral clustering can be attributed to spectral graph theory, where it is used to identify clusters of nodes

within a graph [29]. A basic overview of the spectral clustering algorithm and an example are given below. This paper will build upon these steps to show how the eigengap heuristic can be used to find the optimal number of clusters across multiple scales within a dataset.

For this example, a dataset consisting of 500 two-dimensional data points ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{500}$) was generated as described in the methods (Fig. 1A).

The steps involved in spectral clustering are as follows.

- 1) **An affinity (similarity) matrix, A , is constructed where highly similar (or dissimilar) data points have a high (or low) similarity score.** The Euclidean distance is commonly used to calculate a similarity score. However, several other metrics such as the Manhattan distance [30], Mahalanobis distance [31], Minkowski distance [32], and the Isochronous Temporal metric [33], could be used.

One issue that arises when determining the similarity matrix using a Euclidean metric is that the relationship between similarity and distance is inverted and highly similar points in n -dimensional space have smaller distances than less similar points. To address this issue, the Radial Basis Function (RBF) kernel is used with the Euclidean distance to assign low-valued (or high-valued) distances, to high-valued (or low-valued) similarity scores. The RBF kernel is given by

$$A(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (1)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are data points represented by n -dimensional vectors, $\|\cdot\|^2$ is the Euclidean distance, and the parameter σ determines the width of the kernel. A sorted affinity matrix for this dataset using $\sigma = \sqrt{2}/2$ is shown in Fig. 1C. This value of σ corresponds to the default setting in Python's scikit-learn package [34]. Note that Python's scikit-learn package uses a slightly different version of the RBF kernel where the scaling parameter $\gamma = 1/2\sigma^2$.

- 2) **The normalized Laplacian matrix, L_{sym} , is constructed.** Let $D_i = \sum_{j=1}^n A_{ij}$ denote the degree of node i . The unnormalized graph Laplacian matrix is defined as

$$L = D - A. \quad (2)$$

This matrix is at the core of spectral clustering. A summary of its properties can be found in [35]. The normalized Laplacian matrix used here is $L_{\text{sym}} = D^{-1/2}LD^{-1/2}$. Another option for the normalized Laplacian matrix that is closely related to a random walk is $L_{\text{rw}} = D^{-1}L$ [10].

- 3) **The optimal number of clusters is determined using the eigengap heuristic.** The eigenvalue decomposition of L_{sym} is computed and the resulting eigenvalues are sorted by magnitude. An eigengap is the difference between two successive, sorted eigenvalues. The eigengap heuristic suggests that the number of clusters

should be the number of eigenvalues before the largest eigengap. This represents a 'natural break' within the data, i.e., a region with a stable number of clusters not affected by perturbation.

When eigenvalues of the normalized Laplacian for this dataset are plotted in ascending order as shown in Fig. 1B, the largest eigengap is between the 4th and 5th eigenvalues. Hence, using the eigengap heuristic, four clusters are suggested.

- 4) **The clusters are determined.** Let k be the optimal number of clusters determined in the previous step using the eigengap heuristic. The first k eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ of L_{sym} are stacked into the columns of a matrix. This lower-dimensional matrix is clustered using an aforementioned algorithm, such as the K-means algorithm. The resulting clusters for this dataset are plotted in Fig. 1A and labeled within the affinity matrix in Fig. 1C.

In general, spectral clustering results can be quite sensitive to the choice of σ . A closer look at Fig. 1A and Fig. 1C suggests the possibility of obtaining many more clusters if finer partitions of the data were employed. The choice of $\sigma = \sqrt{2}/2$ corresponds to the σ employed for implementing spectral clustering using scikit-learn but from the discussion above, it is evident that this option is not necessarily optimal. Different values of σ could lead to differences in both the number of suggested clusters and the structure of the resulting clusters. This is due to the influence of σ on the notion of similarity. If σ is too small, only points extremely close to each other are considered similar, and if it is too large, points further away from each other could be considered similar. Therefore, what is considered "optimal" depends on the goal of clustering and is left to the discretion of the researcher.

Here, the scaling factor σ is leveraged to (1) determine how eigengaps change with respect to σ , leading to multiple persistent clustering schemes across the scales of a dataset, (2) determine the persistence of clustering schemes as σ is varied, and (3) determine the stability of nodes within a given clustering scheme across its associated range of σ . Hopefully, this manuscript will convince the reader that there does not have to be one optimal clustering scheme but rather multiple optimal scale-dependent clustering schemes.

III. METHODS

A. DATASETS

1) TWO-DIMENSIONAL TEST DATASET

This dataset originated from a tree-like network of 500 nodes connected by unweighted edges. The positions of the nodes within the xy -plane were obtained using the ForceAtlas2 layout algorithm [36] in Gephi, a network visualization and analysis software [37]. A full description of the network creation method is given in [38]. In summary, the key parameter in the network construction is the branching parameter, ϕ . During this process, a parent node has child nodes added to it if a pseudo-random number $p \in [0, 1]$ is less

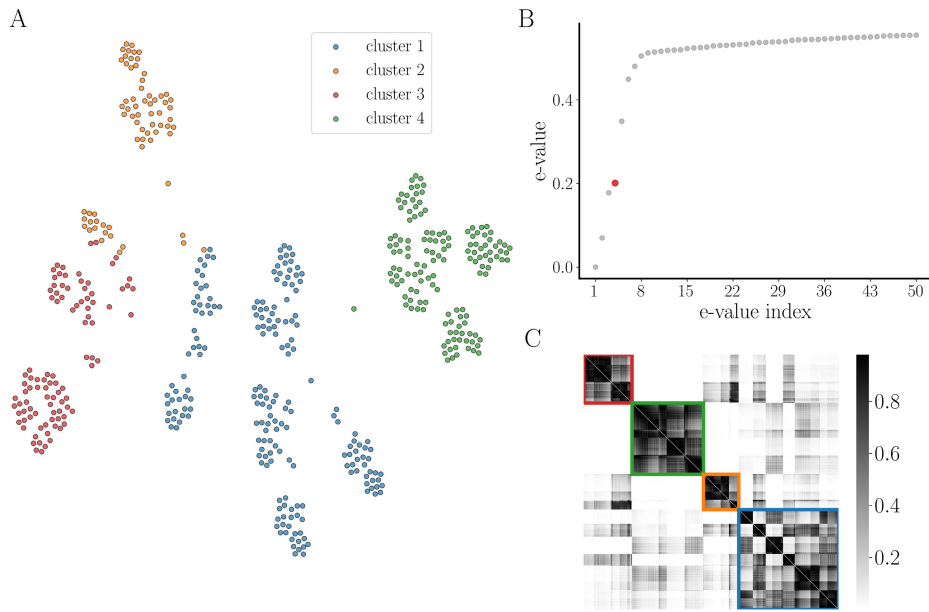


FIGURE 1. Spectral clustering results for the two-dimensional test dataset. Using the RBF kernel with $\sigma = \sqrt{2}/2$, four clusters were suggested. (A) Scatter plot of the two-dimensional test dataset with the four suggested clusters colored red, green, orange, and blue. (B) Plot of sorted eigenvalues of the normalized Laplacian matrix illustrating the index of the largest eigengap. (C) Similarity matrix of the two-dimensional test dataset sorted by cluster, depicting the high similarity within the four clusters, along with possible sub-clusterings.

than ϕ . This generates a rooted tree that stops growing when there are 500 nodes. A high value of ϕ causes the emergence of hub (high degree) nodes in the network, resulting in more discernible clusters. The value of ϕ used here was 0.7.

2) WISCONSIN DIAGNOSTIC BREAST CANCER (WDBC) [39]

The WDBC dataset, obtained from Python's `sklearn.datasets` module, consists of attributes that were computed from digitized images of fine needle aspirates (FNA) of breast masses. It is made up of 569 samples classified as benign (357) or malignant (212), and 10 attributes on each sample. These attributes (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) describe the features of the cell nuclei observed in each image. Each feature in the dataset was preprocessed to have a mean of zero and unit variance using the `StandardScaler` encoding from `scikit-learn`.

3) MODIFIED NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (MNIST) DIGITS DATASET [40]

The Digits dataset was also obtained from Python's `sklearn.datasets` module and contains 1797 samples of 8×8 grayscale pixel images resulting in roughly 180 samples per digit. Each sample represents a hand-written digit image. Each sample contains 64 features (i.e., pixels) that range in value from 0 to 16. Three features with a standard deviation of zero were dropped. No feature scaling was performed

on the data, since each column was already scaled between 0 and 16.

B. METHODS USED WITHIN MSSC

1) CHOICE OF σ RANGE

This set depends on the distribution of distances between points in the given dataset and is chosen such that, for each σ , the resulting affinities are sufficiently large to capture the underlying local and global community structures. While the range of σ is determined by the properties of the dataset, in practice it is recommended to start by defining $\sigma_{\min}$ and $\sigma_{\max}$ using the pairwise Euclidean distances sorted in ascending order: $\sigma_{\min} = P_1(d_{ij})$ and $\sigma_{\max} = P_{60}(d_{ij})$, where $P_k(d_{ij})$ denotes the k -th percentile of pairwise distances. This heuristic generally provides a balance between capturing local and global structures, although adjustments may be necessary for datasets with atypical density, clustering, or outlier distributions. Visual inspection of the distance distribution can help verify that the chosen range effectively spans the relevant structural scales. Within this range, 50 σ values are uniformly sampled to compute the affinities. Adjusting the number of sampled σ values (e.g., sample size = 25 or 100) did not affect the persistent regions (see Supplementary Material for details using the two-dimensional test dataset).

2) INTER-CLUSTER AFFINITY FRACTION

Inter-cluster affinity fraction (IAF_σ). For a given value of the parameter σ and its corresponding set of clusters, this fraction quantifies the sum of affinities between points that belong

to different clusters relative to the sum of affinities between points inside the same cluster.

To compute the IAF_σ , let C_σ represent the set of clusters generated at a particular value of σ . Each point i is assigned to a cluster $\xi_i \in C_\sigma$. Let $A_{i,j}$ denote the affinity between points i and j . The IAF_σ is then defined as

$$IAF_\sigma = \frac{\sum_{\{i,j|\xi_i \neq \xi_j\}} A_{i,j}}{\sum_{\{i,j|\xi_i = \xi_j\}} A_{i,j}} \quad (3)$$

where i and j denote points and the corresponding clusters involved are $\xi_i, \xi_j \in C_\sigma$. A higher IAF_σ indicates more similarity between different clusters, while a lower IAF_σ suggests that within-cluster similarity dominates.

3) NODE CLUSTERING UNCERTAINTY

Although the number of clusters remains fixed within a persistent region, the clustering scheme (i.e., the assignment of points to individual clusters for a given σ) can vary across different values of σ within that region. A node's uncertainty about its cluster membership is measured based on the changes in the set of nodes with which it co-clusters throughout the persistent region. This uncertainty is given by

$$CU_i = 1 - \frac{\sum_{j \in N_i} \kappa_{i,j}}{\sum_j \kappa_{i,j}} \quad (4)$$

where N_i is the set of nodes that always co-cluster with node i , and $\kappa_{i,j}$ is the number of co-clustering incidents between nodes i and j throughout the persistent region. If $CU_i = 0$, then node i always co-clusters with the same set of nodes throughout the persistent region, indicating stable cluster membership. Conversely, if $CU_i = 1$, then node i never co-clusters consistently with any particular set of nodes, suggesting high uncertainty in its cluster assignment. Such nodes are typically located near community boundaries, where membership is more ambiguous.

C. PERFORMANCE METRICS USED WITHIN MSSC

1) SILHOUETTE SCORE

The Silhouette score (SIL) measures how well each data point fits within its assigned cluster compared to other clusters [41], [42]. The score for each point typically ranges between -1 and 1 , where scores closer to 1 suggest that points are well-matched to their cluster and poorly matched to their neighboring clusters, while scores near 0 indicate that points lie on the boundary between two clusters. Negative values may indicate misclassification. The average SIL across all points provides an overall measure of clustering quality. The SIL for a given point is calculated as [34]

$$SIL = \frac{b - a}{\max(a, b)} \quad (5)$$

where a is the mean intra-cluster distance and b is the mean nearest-cluster distance.

2) DAVIES-BOULDIN INDEX

The Davies-Bouldin index (DBI) evaluates the “average similarity” between each cluster x_i , for $i = 1, \dots, k$ and its most similar one x_j , where similarity R_{ij} , is defined as the ratio of within-cluster variation, to between-cluster separation [34]. The minimum score is zero, with lower values indicating better clustering. One way of constructing R_{ij} such that it is nonnegative and symmetric is

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (6)$$

where s_i the average distance between each point of cluster i and the centroid of that cluster, and d_{ij} the distance between cluster centroids i and j .

Hence, the DBI is defined as

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij} \quad (7)$$

3) CALINSKI-HARABASZ INDEX

The Calinski-Harabasz index (CHI), also known as the Variance Ratio Criterion, is used to assess the quality of clustering by examining the ratio of between-cluster dispersion to within-cluster dispersion [34], [43]. A higher CHI generally suggests better clustering quality. The CHI is calculated using the following formula

$$CHI = \frac{BGSS/(k-1)}{WGSS/(n-k)} \quad (8)$$

where BGSS is the between-cluster sum of squares, WGSS is the within-cluster sum of squares, k is the number of clusters, and n is the total number of data points.

4) ADJUSTED RAND INDEX

The Adjusted Rand index (ARI) is an adjustment of the Random Index (RI), which measures the agreement between two clustering assignments, typically the predicted clustering and the ground truth labels [34], [44]. It measures how well the clustering algorithm recovered the actual classes, adjusting for chance agreement. The value ranges from 0 to 1 , where 1 indicates a perfect match between the partitions, and an index close to 0 indicates a random assignment of data points to clusters. The ARI score is computed using the following scheme

$$ARI = \frac{RI - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}} \quad (9)$$

where RI counts the number of pairs of elements that are in either the same group or in different groups in both clusterings, Expected RI is the expected value of RI for random clusterings, and Max RI is the maximum value RI can take.

5) NORMALIZED MUTUAL INFORMATION

The Normalized Mutual Information (NMI) is an external clustering evaluation metric that measures the similarity between two label assignments, such as a clustering result and the ground truth class labels [34]. It ranges from 0 to 1, with higher values indicating greater similarity. It is based on information theory and quantifies how much information is shared between the two clusterings, normalized to account for scale. The NMI is calculated as

$$\text{NMI}(Y, C) = \frac{2I(Y; C)}{\text{mean}(H(Y) + H(C))} \quad (10)$$

where Y represents class labels, C represents cluster labels, $H(\cdot)$ represents Entropy, and $I(Y; C)$ represents Mutual Information between Y and C .

IV. RESULTS

A. SCALE-DEPENDENT CLUSTERINGS WITH MSSC

Local and global graph characteristics are captured by varying σ . The RBF kernel maps small (or large) edge distances to large (or small) affinity values. The way in which the RBF kernel (Eqn 1) transforms edge distances into affinity values is determined by σ . For the two-dimensional test dataset, when σ is small ($\sigma_{\min} = 0.079$; blue curve in Fig. 2) only 1% of edge distances have affinity greater than 0.5 (histogram in Fig. 2A). The corresponding eigenvalue decomposition (blue curve in Fig. 2B) exhibits the largest eigengap before the 34th eigenvalue, suggesting the presence of 33 clusters.

As σ increases, the number of edges with high affinity also increases (green through red curves in Fig. 2A), leading to shifts in the decomposition of the eigenvalues (green through red curves in Fig. 2B). New eigengaps may emerge, while existing ones may collapse as σ varies, resulting in changes to the location of the largest eigengap (black dots in Fig. 2B and Fig. 2C).

The indices of the largest eigengaps appear to decrease monotonically with increasing σ (Fig. 2C). This occurs because a larger σ increases the number of graph edges with high affinity values, effectively creating new connections among the data points. This behavior is reflected in the increase of the inter-cluster affinity fraction (Eq. 3; brown trajectory in Fig. 2C) as σ increases. A sudden drop in the fraction signals the emergence of new clusters.

When there are separations between groups within the data, the cluster structure can remain the same for successive values of σ . If the number of predicted clusters remains the same for a consecutive set of σ values, the clustering is referred to as a “persistent partitioning scheme” and its associated range of σ values is called a “persistent region” (Fig. 2C). The clusters derived from each of these persistent regions are robust, as they are less sensitive to local changes in σ , and may also reflect different scale-dependent attributes of the data.

Graph representations of the two-dimensional test dataset (Fig. 3, second column) illustrate the progression of cluster

partitions for each persistent region shown in Fig. 2C. This progression is also reflected in the heatmap representations (Fig. 3, first column), where sub-blocks are colored according to their corresponding cluster membership.

In region 1, where σ is relatively small, spectral clustering prioritizes local similarities with 13 identified clusters. As σ increases, an increase in the number of inter-cluster edges results in the joining of two clusters that are relatively close in region 1, resulting in 12 clusters (region 2). When σ is relatively large, the number of high-affinity edges increases and brings together individual clusters that are close to one another but relatively further away from other clusters. This leads to a significant increase in the number of inter-cluster edges and a decrease in the number of clusters (regions 3 and 4).

Within a persistent region, as σ increases, certain edges are added to the graph, which could affect the cluster membership of specific nodes. The last column of Fig. 3 shows nodes that changed their cluster membership as σ increased through each persistent region. Points on or near the boundaries of closely situated clusters tend to be less stable as the inter-cluster edges increase (regions 1, 2 and 3). When clusters are fairly distant from each other, cluster membership tends to remain consistent throughout the region (region 4).

These observations are supported by cluster validity indices, which quantify cluster cohesiveness and separation. The SIL values, which range from approximately 0.5 to 0.6 (Fig. 4), and the CHI measures, which range from about 700 to 2100 (Fig. 4), both suggest that the clusters are generally well-separated and compact. The DBI values, which range from approximately 0.55 to 0.80 (Fig. 4), further reinforce the stability of the clusters within the persistent regions. These metrics are consistent with the structure of the dataset, which comprises dense and sparse regions (Fig. 1). Although not exceptionally high, these scores provide evidence of a meaningful underlying clustering structure. Within persistent regions, minor variations in the SIL and CHI values suggest small changes in cluster membership and size; however, these fluctuations remain limited. More pronounced changes are observed at the boundaries between regions, indicating transitional uncertainty. A distinct peak in the SIL curve that coincides with the minimum in the DBI plot at the onset of region 2, highlights a phase of increased cluster compactness and separation. As σ increases and clusters begin to merge, the SIL and CHI values gradually decrease due to the increasing intra-cluster dispersion, although the effect remains relatively small.

B. IDENTIFYING PERSISTENT REGIONS FOR LABEL ALIGNMENT USING MSSC

The previous example highlighted the dependency of the resulting MSSC clusters on scale. To examine how clusters in persistent regions align with known class labels, MSSC was applied to the WDBC dataset, which includes ground truth labels for benign and malignant tumor samples. Using Python’s default RBF kernel setting of $\sigma = \sqrt{2}/2$, the results

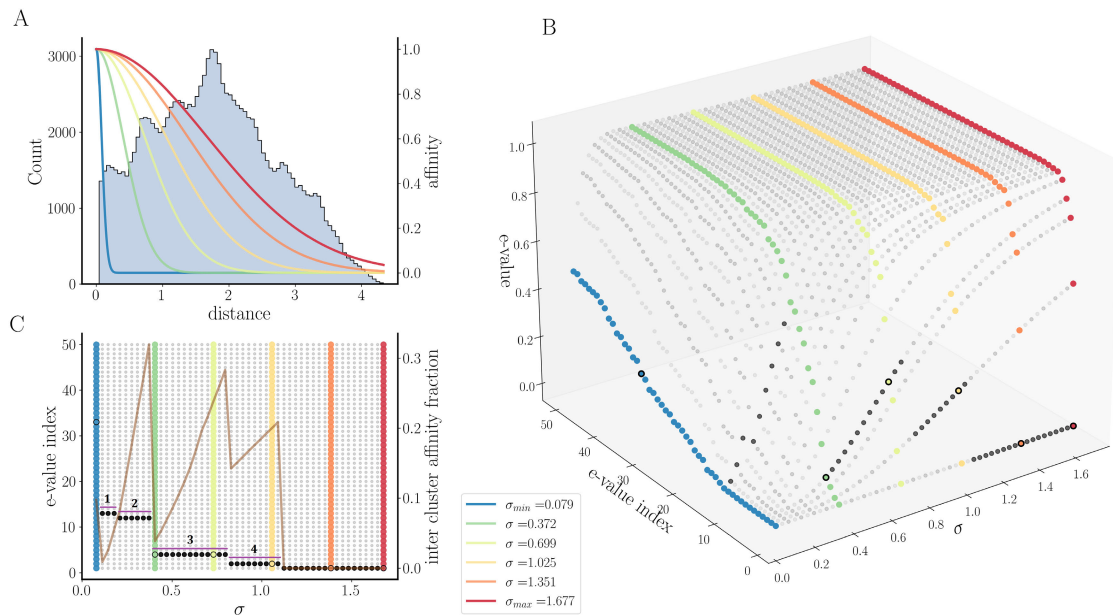


FIGURE 2. The effects of varying σ on the number and composition of clusters of the two-dimensional test dataset. (A) Histogram of distance data superimposed with six affinity curves. (B) Plot of eigenvalues showing the largest eigengaps (black dots) corresponding to the 50 σ values sampled from [0.079, 1.677]. (C) Plot showing four persistent partitioning schemes (not including the case where a single cluster is suggested) of the data superimposed with a trajectory (brown) of the inter-cluster affinity fraction.

of the eigengap heuristic suggests eight clusters (Fig. 5). The block-diagonal heatmap in Fig. 5C first partitions the dataset into the two cancer groups and then into sub-blocks representing the clusters. For this value of σ , three of the clusters contain members of the benign and malignant classes, while five of them contain only malignant tumors. None of the clusters are exclusively made up of benign tumors.

Subsequently, MSSC was employed to assess how the number of clusters varied across different scales (Fig. 6). As with the two-dimensional test dataset, a histogram of pairwise distances overlaid with six affinity curves is constructed (Fig. 6A). The eigenvalue decomposition computed for each of the 50 σ values (Fig. 6B) yields four persistent regions (Fig. 6C).

In region 1, where σ is relatively small, spectral clustering identifies eight clusters (Fig. 7). In particular, most edge distances with affinity greater than 0.5 are concentrated within the olive cluster, which is predominantly composed of benign tumors. This suggests that the benign tumors in the dataset likely exhibit a high degree of similarity. However, this persistent region is transient.

As σ increases, more inter-cluster edges form, resulting in the merger of three nearby clusters (pink, brown, and red in region 1) into a single red cluster in region 2 (Fig. 7). Meanwhile, the brown cluster in region 1 splits, with some members joining the olive cluster to form a new purple cluster in region 2, while others join the newly formed red cluster. These changes lead to a reduction in the number of clusters to five in region 2. This region is also transient.

Note that even though these two regions meet the formal definition of persistence, they are relatively short-lived and may not be as robust as other regions. Choosing a σ value within these regions could lead to results that are highly sensitive to minor variations in the data structure.

As σ increases further, the eigenvalue decomposition changes again. The previously merged cluster in region 2 separates into its original groups, while the previously split clusters reassemble. This restores the eight-cluster configuration observed in region 1 and defines region 3.

At relatively higher values of σ , high-affinity edges increase, causing the nearby clusters to merge, resulting in the formation of four clusters. This configuration remains stable for a considerably longer period. Region 4 persists from $\sigma = 0.729$ to 1.175, beyond which only one cluster is suggested.

The heatmaps in Fig. 7 suggest that the spectral clustering results for regions 2 and 4 in particular were very close to the ground truth labels. However, the clusters depicted in the four persistent regions may reflect some characteristics of the dataset that are not included in the metadata.

To better understand the implications of the stability regions, cluster quality was evaluated using both external and internal validation metrics. The ARI and NMI values (Fig. 8A) generally indicate a strong alignment between the clustering results and the two cancer classes. This correspondence is particularly evident in regions 2 and 4, where peak ARI and NMI values are observed at the onset of region 4. These findings are consistent with the observation that the results of spectral clustering in these regions closely

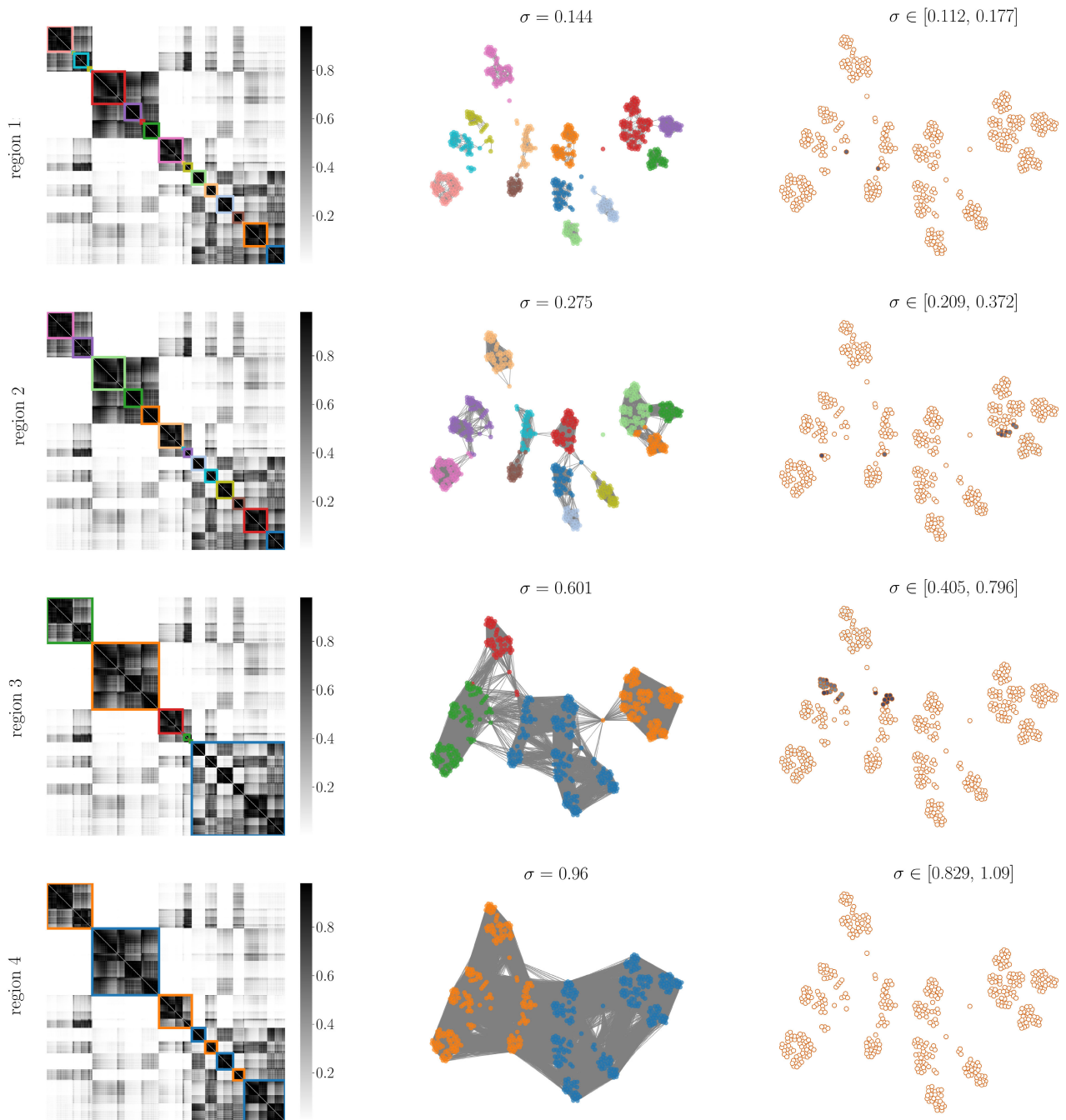


FIGURE 3. Analysis of clusters corresponding to the persistent regions in Fig. 2C. Persistent regions 1, 2, 3, and correspond to 13, 12, 4, and 2 clusters, respectively. The first column depicts sub-blocks (representing clusters) corresponding to each persistent region, superimposed on the block-diagonal heatmaps representing the four clusters shown in Fig. 1. Scatter plots of the clusters along with edges between points are depicted in the second column. Edges denote node pairs with affinity values greater than 0.5. They are included for illustrative purposes only. The third column shows scatter plots of points along with brown dots that represent points that had the propensity to belong to other clusters.

match the ground truth labels. In contrast, the SIL and CHI values are generally lower, while the DBI scores are higher compared to those observed for the test dataset (Fig. 8B). This is due to the close proximity and potential overlap of the clusters in the WDBC dataset. Moreover, since some of the clusters generated for the WDBC dataset are non-convex or

irregularly-shaped, the SIL and CHI may produce misleading results when applied [45]. However, relative to regions 1 and 3, regions 2 and 4 exhibit higher SIL and CHI values, despite having fewer clusters, suggesting that merged clusters in these regions are more distinct and less overlapping. As observed in the test dataset, sharp fluctuations in the SIL

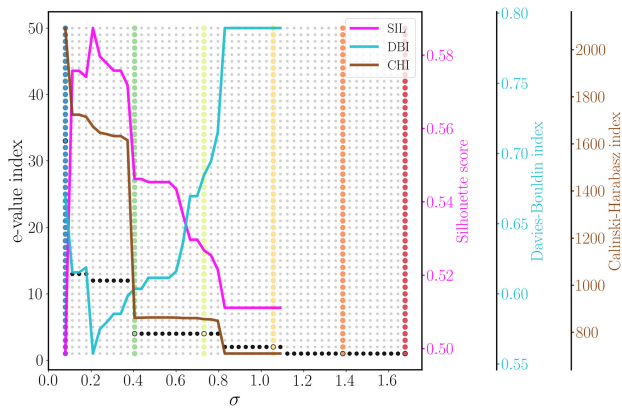


FIGURE 4. Performance metrics for clustering outcomes on the two-dimensional test dataset across varying σ values. SIL values ranging from approximately 0.5 to 0.6, and CHI values lying between about 700 and 2100, suggest well-separated and compact clusters. DBI values, ranging from approximately 0.55 to 0.80 further support the robustness of clusters within the persistent regions.

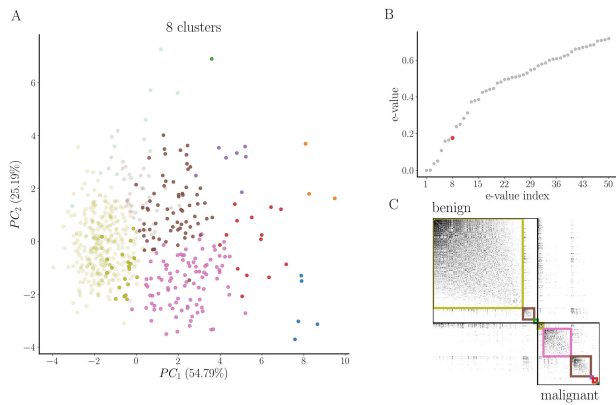


FIGURE 5. Spectral clustering results for the WDBC dataset. Eight clusters were generated when the RBF kernel with $\sigma = \sqrt{2}/2$ was used. (A) Scatter plot of the eight clusters: olive, green, pink, brown, orange, red, purple, and blue. Solid and faded points denote malignant and benign tumors, respectively. (B) Plot of the eigenvalues showing the index of the largest eigengap. (C) Block-diagonal heatmap of the two breast cancer classes superimposed with sub-blocks corresponding to the eight clusters.

and CHI values at the boundaries of the persistent regions indicate increased uncertainty in cluster assignments during transitions.

C. VISUALIZATION OF SCALE-DEPENDENT CLUSTERINGS

While the WDBC dataset has existing labels, it is not easy to visualize how cells differ between clusters. The Digits dataset gives us a unique opportunity to visualize differences between scale-dependent clusters. For the Digits dataset, a Uniform Manifold Approximation and Projection (UMAP) [46], was used to visualize the class distribution of the data (Fig. 9). Some digit classes (such as the twos, fours, and sixes) form clearly well-separated clusters within the UMAP projection, whereas other digit classes (such as the ones and nines) do not form a single cohesive cluster. MSSC was applied to the Digits dataset (Fig. 10, note that UMAP was used for

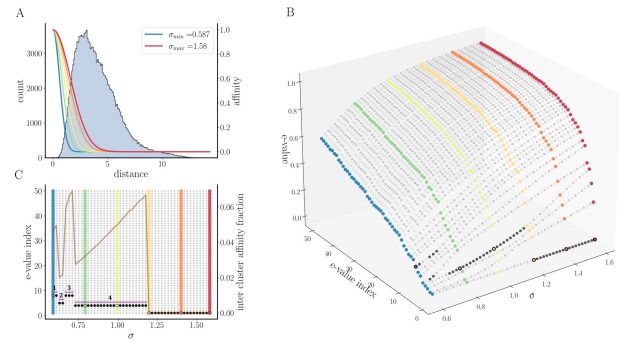


FIGURE 6. The effects of varying σ on the number and composition of clusters of the WDBC dataset. (A) Histogram of distance data superimposed with six affinity curves. (B) Plot of eigenvalues showing the largest eigengaps (black dots) corresponding to the 50 σ values sampled from [0.587, 1.58]. (C) Plot showing various partitioning schemes of the data superimposed with a trajectory (brown) of the inter-cluster affinity fraction. If the set [0.587, 1.580] is partitioned into five equal parts, then the rest of the affinity curves (from left to right) correspond to a σ value of 0.79, 0.98, 1.18, and 1.38, respectively.

visualization, not as a preprocessing tool) to group similar patterns of handwritten digits. For this dataset, $\sigma_{\min}$ and $\sigma_{\max}$ were found to be 4.494 and 28.436, respectively. The values of σ are much higher than those observed in the previous datasets, since the range for each feature was [0, 16].

Based on the largest eigengaps (Fig. 10B), three persistent regions were found, corresponding to 22, 12, and 10 clusters, respectively (Fig. 10C). The spectral clustering results corresponding to each of the three persistent regions are depicted in Fig. 11. Every heatmap in the first column displays 10 diagonal blocks, each representing a digit class in the dataset, along with sub-blocks that show the clusters within each class. The composition of digits within each cluster is calculated, together with the average images of different digit types in each cluster (Fig. 12).

In region 1, 11 of the 22 clusters are digit-specific, with 100% of their members belonging to a single digit class (top row in Fig. 12). Of the remaining 11 clusters, nine have more than 90% representation of a single digit, one has 86%, and the other has 60%. The zero digit class (deep brown, encircled in Fig. 12) forms a unique, tightly clustered group, indicating that it shows little variation within the dataset. The pink and purple clusters (hexagon-shaped in Fig. 12) are composed almost exclusively of distinct digits, namely, three and six, respectively, with only a few images clustered among other groups. The fives digit class appears to be clustered into two groups (orange and pinkish purple, right-triangle-shaped in Fig. 12), depicting two patterns of the same digit. Similarly, the ones digit class is divided into four subclasses (light blue, pinkish orange, olive, and light green, as boxed in Fig. 12), suggesting a significant variation in the way those digits were written.

However, in region 2, where σ increases, some clusters from region 1 that were previously adjacent merge into a single cluster. For example, the pink cluster (mainly composed of threes) and the brown cluster (mainly composed of nines) in region 1 are now grouped into the blue cluster in

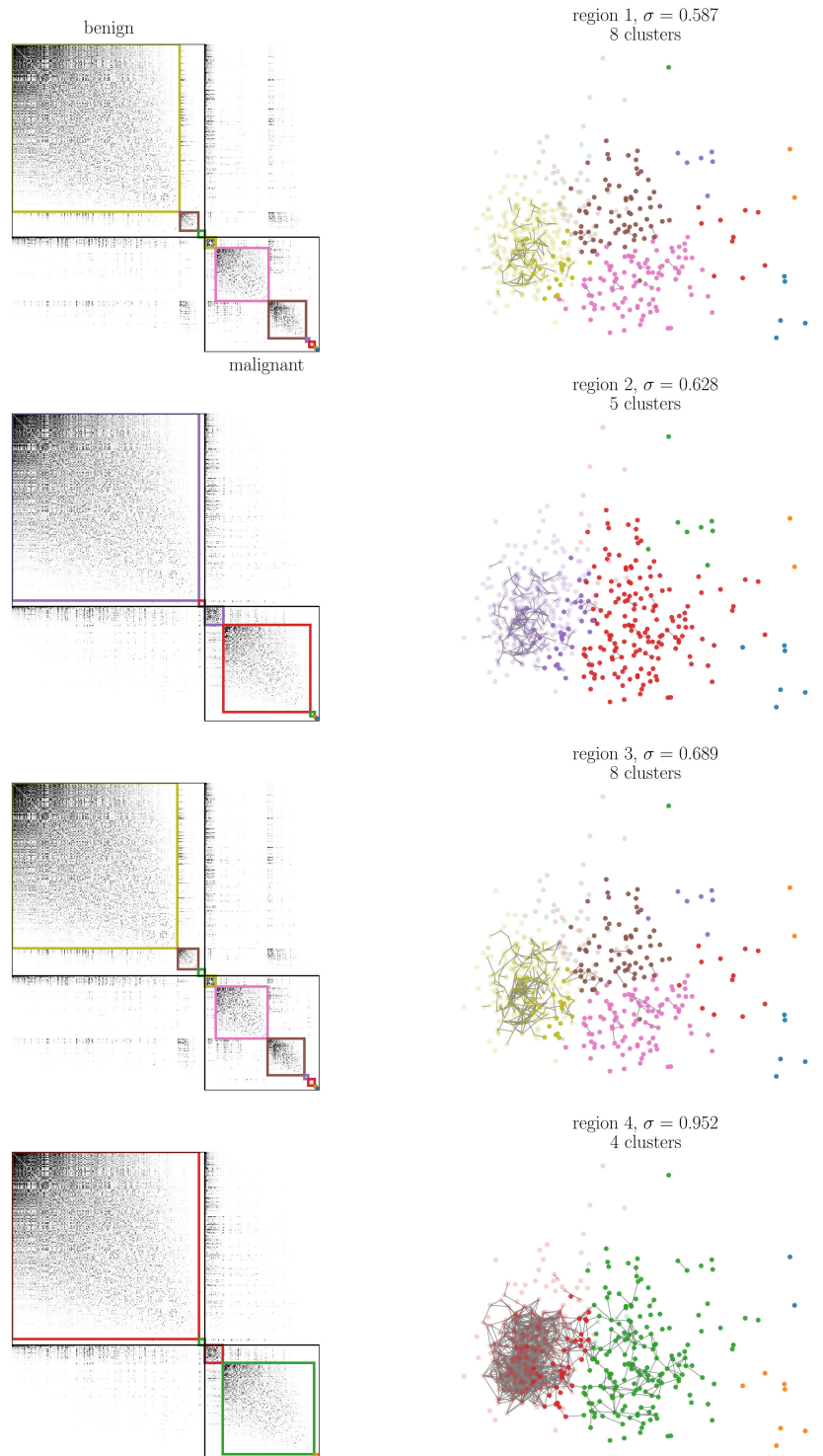


FIGURE 7. Analysis of clusters corresponding to the persistent regions in Fig. 6C. Persistent regions 1, 2, 3, and 4 correspond to 8, 5, 8, and 4 clusters, respectively. The first column depicts block-diagonal heatmaps of the clusters corresponding to each persistent region. Each heatmap contains two blocks describing the two classes of cancer as well as sub-blocks describing individual clusters within a class. A scatter plot of the clusters together with edges between points are depicted in the second column. Edges denote node pairs with affinity values greater than 0.5.

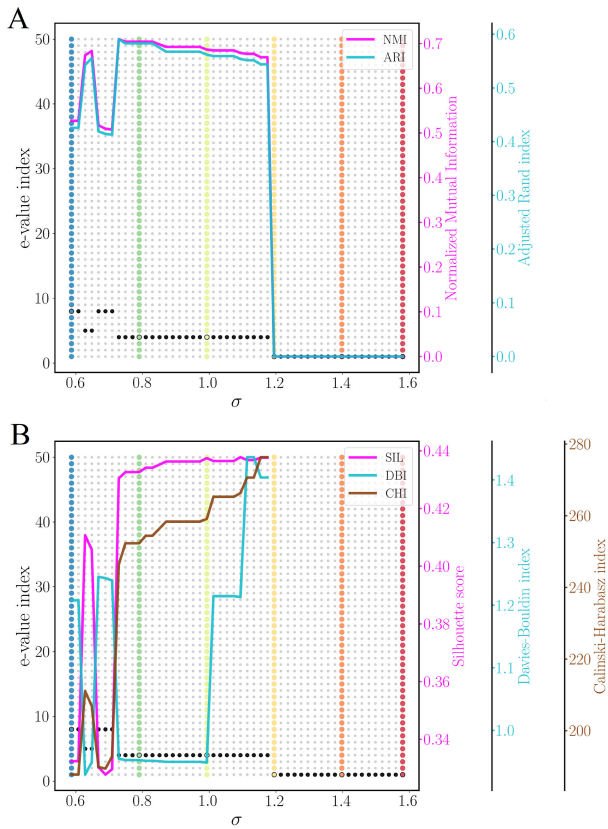


FIGURE 8. Performance metrics for clustering outcomes of the WDBC dataset across varying σ values. (A) The ARI values ranging from approximately 0.42 to 0.60, and the NMI values lying between about 0.50 and 0.70, indicate a strong correspondence between the clustering results and the two cancer classes. (B) The SIL values ranging between about 0.32 and 0.44, the CHI values ranging between approximately 180 and 280, and the DBI values lying between about 0.9 and 1.5, suggest close proximity among clusters. Importantly, these values do not imply poor clustering performance but rather reflect the inherent structure and cluster proximity within the WDBC dataset.

region 2 (Fig. 12). However, the olive cluster (mainly made up of fours), which was split into two separate clusters in region 1, is now combined into a single cluster in region 2 (Fig. 12).

Region 3 further simplifies the clustering by merging additional groups (e.g., the red and orange clusters in region 2, Fig. 12). This results in a more generalized classification with less specificity in some clusters than in region 2.

The comparison of the clustering schemes for regions 1, 2, and 3 reveals key differences in the specificity of the members and the structure of the clusters. In region 1, the data are divided into a greater number of distinct clusters, suggesting a more precise classification that captures subtle variations among the images. Moreover, the right-hand legends in Fig. 12 provide representative images for each cluster, suggesting that the clusters in region 1 are more specific, with fewer mixed categories within a single cluster. In contrast, regions 2 and 3 consolidate some of these clusters, exhibiting more variability within some clusters, a broader range of data representations, and a reduction in the total number of clusters

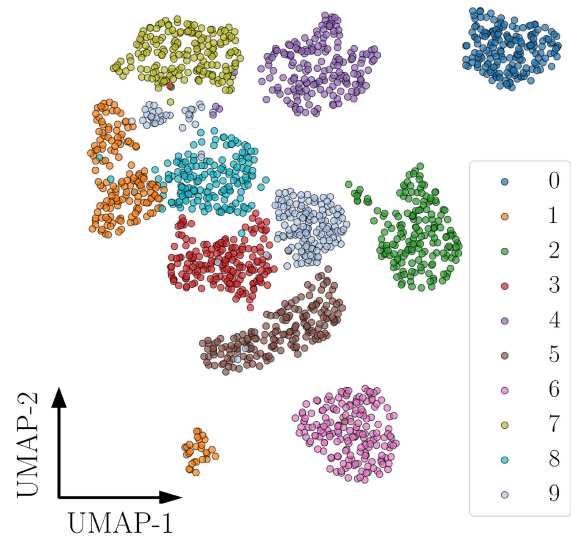


FIGURE 9. A UMAP representation of the MNIST digits dataset.

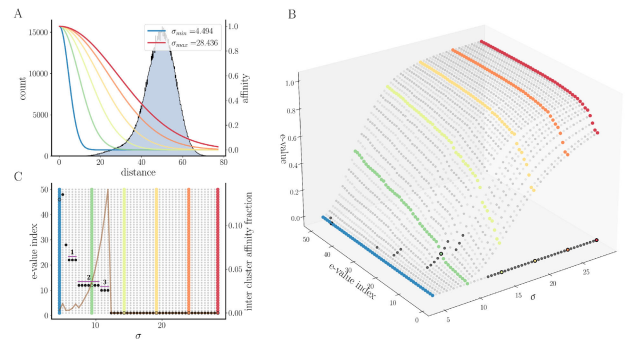


FIGURE 10. The effects of varying σ on the number and composition of clusters of the MNIST digits dataset. (A) Histogram of distance data superimposed with six affinity curves. ($\sigma_{\min}$ and $\sigma_{\max}$) were found to be 4.494 and 28.436, respectively with the four other affinity curves (from left to right) corresponding to σ values of 9.28, 14.07, 18.86, and 23.65, respectively. (B) Plot of eigenvalues showing the largest eigengaps (black dots) corresponding to the 50 σ values sampled from [4.494, 28.436]. (C) Plot showing various partitioning schemes of the data superimposed with a trajectory (brown) of the inter-cluster affinity fraction.

to 12 and 10, respectively. This indicates a more generalized grouping, with some clusters that were previously distinct in region 1 appearing to have been merged in regions 2 and 3, resulting in broader but potentially less homogeneous groupings. The visualization also highlights an isolated cluster in the upper-right corner of all rows in Fig. 12, though its composition differs slightly between the three regions. These differences suggest that persistent regions 2 and 3 prioritize broader generalization over specificity, whereas region 1 emphasizes finer differentiation. Therefore, the choice among these three clustering schemes depends on the specific research objectives, whether the focus is on achieving higher precision in differentiation (region 1) or optimizing for a more generalized categorization (regions 2 and 3).

Finally, to assess clustering quality and compare the results with the known digit classes, the performance metrics discussed in the previous example were applied. The ARI and NMI values were generally high, indicating a strong

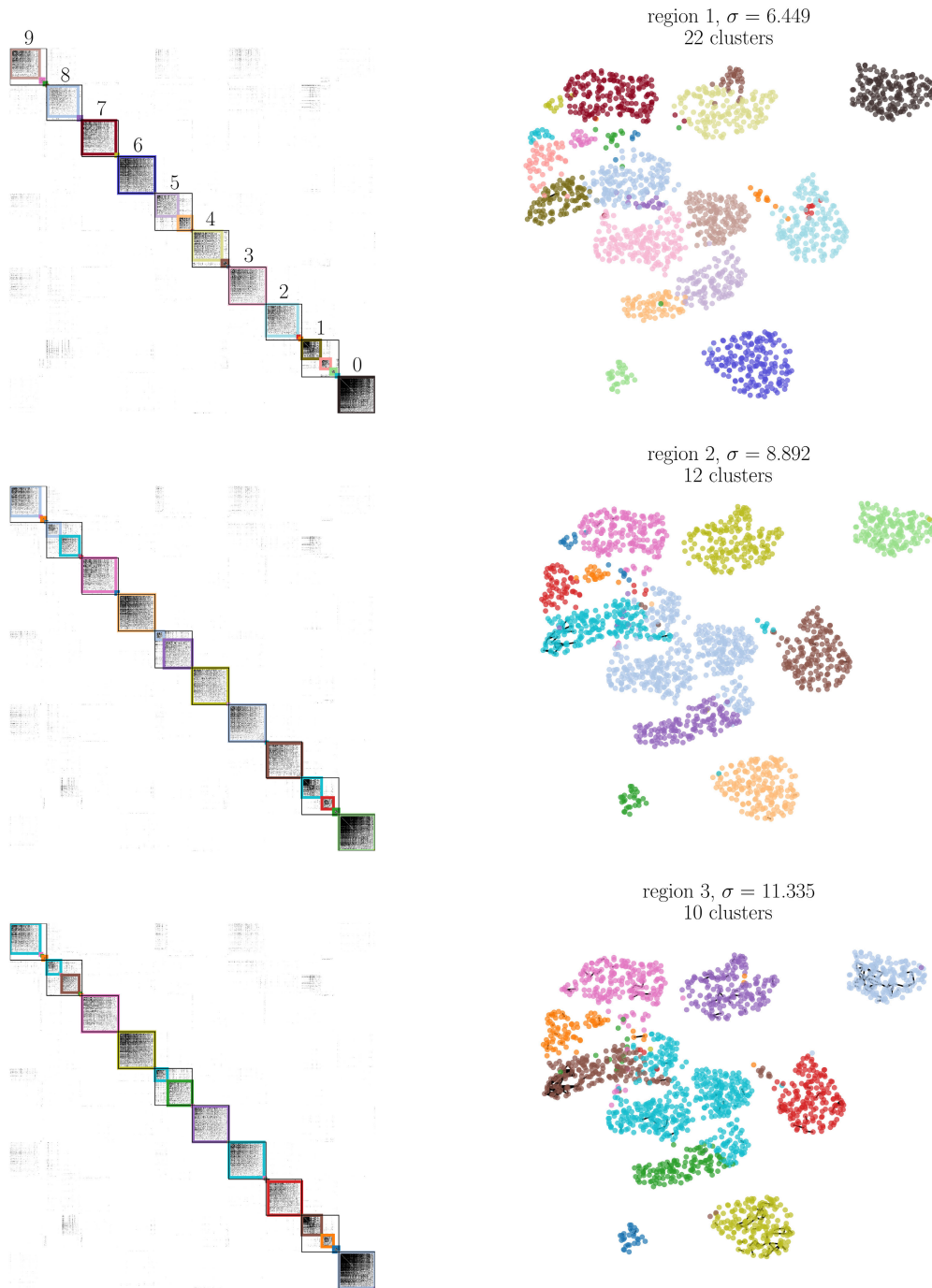


FIGURE 11. Analysis of clusters corresponding to the persistent regions in Fig. 10C. Persistent regions 1, 2, and 3 correspond to 22, 12, and 10 clusters, respectively. The first column shows block-diagonal heatmaps of the clusters corresponding to each persistent region. Each heatmap contains 10 blocks describing the 10 digit classes as well as sub-blocks describing individual clusters within a class. Scatter plots of the clusters together with edges between points are depicted in the second column.

correspondence between the clustering outcomes and the ten digit classes (Fig. 13A). This agreement peaked in region 1, where clusters exhibited greater specificity with respect to digit identity. In contrast, the SIL and CHI values were

low, while the DBI scores were high compared to those observed in the two previous examples (Fig. 8B). This is attributed to the close proximity and significant overlap among clusters within the persistent regions, largely due to

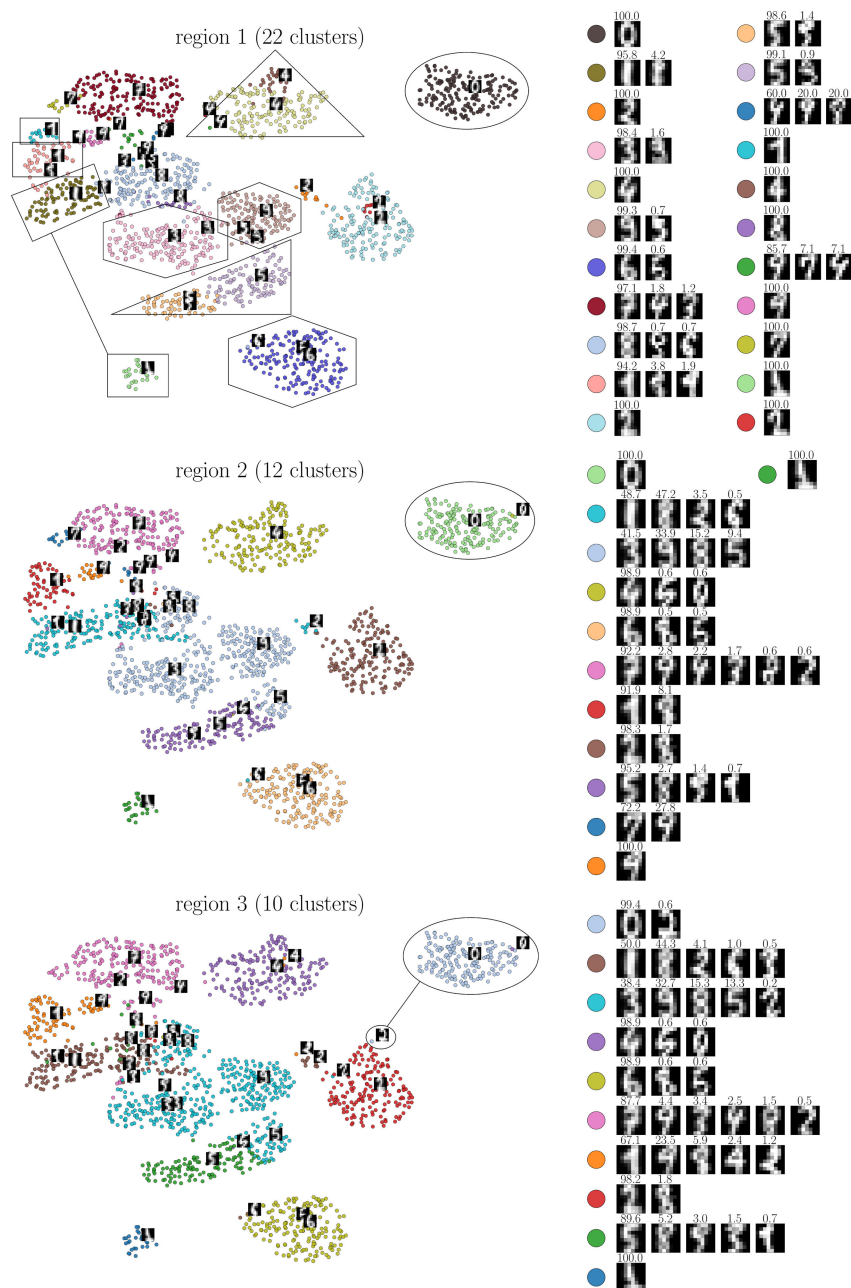


FIGURE 12. Further analysis of the clusters corresponding to the persistent regions in Fig. 10C. Each cluster presents the average shapes of handwritten digit images. The average shapes are depicted using grayscale images, while the percentage of each shape within a cluster is displayed in the legend on the right. The representation also identifies an isolated cluster (encircled) in the upper-right corner across all rows with slight variations in its composition among the three regions.

the visual similarity between certain digits. This effect was particularly evident in region 1, which contained 22 clusters, many of which were close to each other or exhibited overlap. However, as with the WDBC dataset, increasing the value of σ led to less homogeneous clusters. Consequently, the SIL and CHI values gradually increased due to reduced intra-cluster dispersion. As specificity decreased, digits that were previously distinguished began to merge into broader

categories. This transition resulted in comparatively less overlap, as the clusters became more generalized and better separated.

V. DISCUSSION

Here, a new method is introduced, MSSC, which builds on the principles of spectral clustering to incorporate scalar dependency. This approach illustrates that there may not be

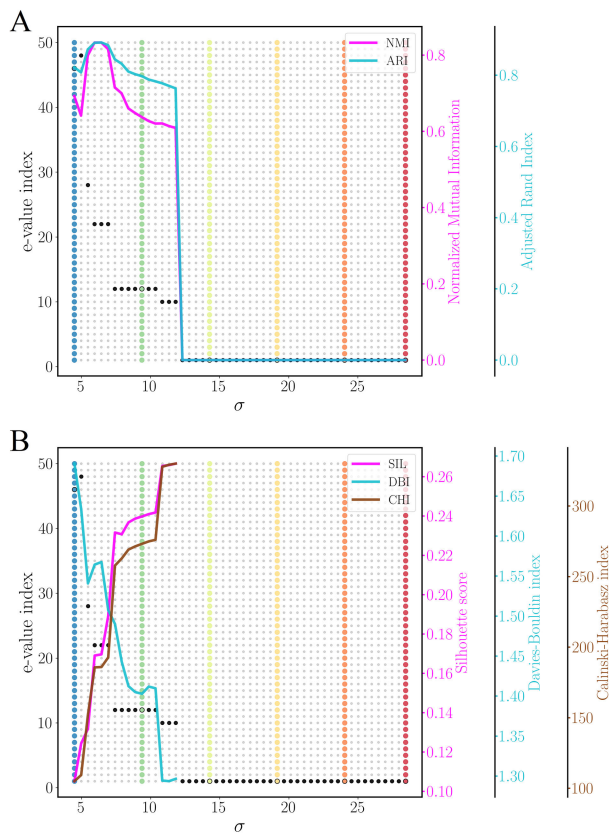


FIGURE 13. Performance metrics for clustering outcomes of the Digits dataset across varying σ values. (A) The ARI values, ranging from approximately 0.8 to 0.9, and the NMI scores, ranging from about 0.6 to 0.8, demonstrate a strong correspondence between the clustering results and the 10 digit classes. (B) The SIL values, ranging between 0.10 and 0.26, CHI values, ranging from approximately 100 to 350, and DBI values, lying between about 1.3 and 1.7, suggest that the identified clusters are not well-separated from one another. However, these values do not imply poor clustering performance; rather, they reflect the intrinsic structure of the Digits dataset, which exhibits several closely situated and partially overlapping clusters within persistent regions.

a single optimal clustering solution for a dataset. Rather, it involves exploring several plausible solutions, each unveiling a different scalar view of the data. This phenomenon is observed in common daily experiences, such as examining an electronic map where regions can be identified based on country, state, or city lines depending on the scale. This scale-dependent notion is also analogous to adjusting the zoom on a microscope and observing different layers of patterns emerge, each dependent on a particular zoom level.

As part of a doctoral dissertation, the MSSC algorithm developed in this work was applied to identify subgroups of septic patients based on their temporal protein biomarker profiles. In that study, the algorithm was integrated with ordinary differential equation (ODE) modeling to characterize immune response dynamics across distinct patient classes. This prior application demonstrated the utility of the method in uncovering clinically relevant immune response patterns, providing an early validation of its potential for analyzing

heterogeneous biomedical data. For more details on the application of the algorithm to sepsis patient data, refer to the dissertation [47].

This novel algorithm is purely data-driven, which allows for direct identification of distinct scale-dependent clusters within a dataset. The scale-dependent number of clusters is determined by the eigengap heuristic, a completely data-driven approach, instead of relying on the elbow of the AIC or BIC curves. Continuous scale values that consistently predict the same number of clusters called persistent regions are introduced, providing a more nuanced understanding of clustering stability across different scales and allowing researchers to assess the robustness of clustering results while identifying meaningful structures within a dataset. The clusters themselves are defined using the K-means algorithm on the dimensionally-reduced similarity matrix. Since K-means is sensitive to initialization, it can yield different results across runs, potentially introducing some variability regarding the clustering scheme. Hence, exploring methods to mitigate this variability could enhance the robustness of this approach.

The datasets to which the algorithm was applied were chosen to demonstrate its adaptability and versatility across different domains. The two-dimensional representation of a tree-like network evaluates its ability to capture hierarchical relationships. The Digits dataset tests the algorithm's effectiveness on an unstructured, discrete, high-dimensional, image-based data with complex clusters, while the WDBC dataset, which is structured and contains continuous data, provides a real-world application in clustering tumors based on physical characteristics.

While some multi-level approaches such as ensemble clustering improve robustness through consensus across multiple results, this approach results in one clustering. MSSC may be preferred when scale sensitivity and stability across scales are informative. However, a new approach could combine the ensemble clustering method with MSSC's clusterings, weighted by a persistence measure, producing one clustering scheme that takes into account the observed persistent regions.

However, the proposed approach could be computationally expensive for very large or high-dimensional datasets, as the method requires eigenvalue decompositions across many scale values. Care must also be taken when choosing bounds, since meaningful structures could be missed. The method currently assumes numerical data and Euclidean distance. A future extension of the algorithm could be to increase its applicability to datasets containing non-numeric or mixed-type features. The Hamming distance [48] can be applied to categorical datasets, while the Gower's distance [49] may be employed for mixed-type data. Lastly, sensitivity to noise and outliers, a common issue in kernel-based clustering methods, can affect the reliability of similarity scores and, consequently, the cluster quality. The isochronous temporal metric [33] could be particularly useful for handling noisy data.

VI. CONCLUSION

In conclusion, the MSSC method offers a novel data-driven approach to identifying scale-dependent clusters, providing deeper insight into the structure and stability of clustering solutions across different scales. By utilizing the eigengap heuristic and introducing the concept of persistent regions, this approach avoids reliance on arbitrary parameter choices and instead highlights robust clustering configurations that consistently emerge within the bounded range of possible clusterings. Rather than defining a single optimal number of clusters, our method identifies robust, data-driven solutions that persist across scales and are less sensitive to noise or parameter variation.

This multiscale perspective has practical value in applications where both local and global data structure matter, such as identifying patient subtypes in biomedical data, detecting communities in networks, or analyzing image and signal patterns at different resolutions. By revealing clustering patterns that remain stable across a range of scales, the method supports more interpretable and reliable results for decision-making, diagnostics, and exploratory data analysis. Future extensions could apply this framework to time-evolving data, integrate automatic scale range selection, or tailor the approach to specific domains such as dynamical systems, genomics, or environmental monitoring, where multi-resolution structure plays a critical role.

ACKNOWLEDGMENT

The authors would like to thank Pavol Genzor, Josh Chenoweth, and Gary Fogel for their advice and support. The views expressed in this manuscript reflect the views of them and do not necessarily reflect the position of The Henry M. Jackson Foundation for the Advancement of Military Medicine Inc., the Department of the Army, the Department of Defense, nor the United States Government. References to non-federal entities or their products do not constitute or imply the Department of Defense or Army endorsement of any company, product, or organization.

REFERENCES

- [1] S. Zeng, R. Huang, Z. Kang, and N. Sang, "Image segmentation using spectral clustering of Gaussian mixture models," *Neurocomputing*, vol. 144, pp. 346–356, Nov. 2014.
- [2] W. Pentney and M. Meilă, "Spectral clustering of biological sequence data," in *Proc. AAAI*, 2005, pp. 845–850.
- [3] K. Berahmand, E. Nasiri, R. P. Mohammadiani, and Y. Li, "Spectral clustering on protein–protein interaction networks via constructing affinity matrix using attributed graph embedding," *Comput. Biol. Med.*, vol. 138, Nov. 2021, Art. no. 104933.
- [4] G. Obadi, P. Dráždilová, J. Martinović, K. Slaninová, and V. Šnášel, "Using spectral clustering for finding students' patterns of behavior in social networks," in *Proc. DATESO*, 2010, pp. 118–130.
- [5] J. G. Chenoweth, J. Brandsma, D. A. Striegel, P. Genzor, E. Chiyka, P. W. Blair, S. Krishnan, E. Dogbe, I. Boakye, G. B. Fogel, E. L. Tsalik, C. W. Woods, A. Owusu-Ofori, C. Oppong, G. Oduro, T. Vantha, A. G. Letizia, C. G. Beckett, K. L. Schully, and D. V. Clark, "Sepsis endotypes identified by host gene expression across global cohorts," *Commun. Med.*, vol. 4, no. 1, p. 120, Jun. 2024.
- [6] S. K. Uppada, "Centroid based clustering algorithms—A clarion study," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, pp. 7309–7313, 2014.
- [7] H. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 231–240, 2011.
- [8] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognit.*, vol. 45, no. 11, pp. 3950–3961, Nov. 2012.
- [9] Z. Liu and M. Barahona, "Graph-based data clustering via multiscale community detection," *Appl. Netw. Sci.*, vol. 5, no. 1, pp. 1–20, Dec. 2020.
- [10] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.
- [11] D. Verma and M. Meila, "A comparison of spectral clustering algorithms," University Washington, Seattle, WA, USA, Tech. Rep. UWCSE030501, 2003, vol. 1, pp. 1–18.
- [12] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability," *Artif. Intell.*, vol. 273, pp. 37–55, Aug. 2019.
- [13] R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha, "A survey: Clustering ensembles techniques," *World Acad. Sci.*, vol. 3, no. 2, pp. 365–374, 2009.
- [14] B. Kılıç and S. Özarpaç, "Ensemble clustering in GPS velocities: A case study of Turkey," *Appl. Sci.*, vol. 12, no. 24, p. 12636, Dec. 2022.
- [15] H. Bozdoğan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, Sep. 1987.
- [16] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *WIREs Comput. Statist.*, vol. 4, no. 2, pp. 199–203, Mar. 2012.
- [17] M. Herle, N. Micali, M. Abdulkadir, R. Loos, R. Bryant-Waugh, C. Hübel, C. M. Bulik, and B. L. De Stavola, "Identifying typical trajectories in longitudinal data: Modelling strategies and interpretations," *Eur. J. Epidemiol.*, vol. 35, no. 3, pp. 205–222, Mar. 2020.
- [18] F. Liu and Y. Deng, "Determine the number of unknown targets in open world based on elbow method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, May 2021.
- [19] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 849–856.
- [20] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 1601–1608.
- [21] G. Sanguinetti, J. Laidler, and N. D. Lawrence, "Automatic determination of the number of clusters using spectral algorithms," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, Sep. 2005, pp. 55–60.
- [22] A. Azran and Z. Ghahramani, "Spectral methods for automatic multiscale data clustering," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 190–197.
- [23] A. Azran and Z. Ghahramani, "A new approach to data driven clustering," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 57–64.
- [24] A. Little and A. Byrd, "A multiscale spectral method for learning number of clusters," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 457–460.
- [25] M. Afzalan and F. Jazizadeh, "An automated spectral clustering for multi-scale data," *Neurocomputing*, vol. 347, pp. 94–108, Jun. 2019.
- [26] C. Zhang, G. Zhu, B. Lian, M. Chen, H. Chen, and C. Wu, "Image segmentation based on multiscale fast spectral clustering," *Multimedia Tools Appl.*, vol. 80, no. 16, pp. 24969–24994, Jul. 2021.
- [27] X. Li and Z. Tian, "Multiscale stochastic hierarchical image segmentation by spectral clustering," *Sci. China Ser. F, Inf. Sci.*, vol. 50, no. 2, pp. 198–211, Apr. 2007.
- [28] X. Wang, Y. Zhang, and Y. Zhou, "Multimodal remote sensing image clustering with multiscale spectral–spatial anchor graphs," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4405612, doi: 10.1109/TGRS.2025.3545460.
- [29] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Comput. Appl.*, vol. 24, nos. 7–8, pp. 1477–1486, Jun. 2014.
- [30] M. D. Malkauthekar, "Analysis of Euclidean distance and Manhattan distance measure in face recognition," in *Proc. 3rd Int. Conf. Comput. Intell. Inf. Technol. (CIIT)*, Oct. 2013, pp. 503–507.
- [31] P. O. Brown, M. C. Chiang, S. Guo, Y. Jin, C. K. Leung, E. L. Murray, A. G. M. Pazdor, and A. Cuzzocrea, "Mahalanobis distance based K-means clustering," in *Proc. Int. Conf. Big Data Anal. Knowl. Discovery*, 2022, pp. 256–262.

- [32] A. A. Thant and S. M. Aye, "Euclidean, Manhattan and Minkowski distance methods for clustering algorithms," *Int. J. Sci. Res. Sci., Eng. Technol.*, vol. 7, no. 3, pp. 553–559, Jun. 2020.
- [33] A. K. Kumar, N. N. Mai, K. Tian, and Y. Xia, "Isochronous temporal metric for neighbourhood analysis in classification tasks," *Social Netw. Comput. Sci.*, vol. 4, no. 6, p. 807, Oct. 2023.
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
- [35] K. C. Das, "The Laplacian spectrum of a graph," *Comput. Math. With Appl.*, vol. 48, nos. 5–6, pp. 715–724, Sep. 2004.
- [36] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLoS ONE*, vol. 9, no. 6, Jun. 2014, Art. no. e98679.
- [37] M. Bastian, S. Heymann, and M. Jacomy. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. [Online]. Available: <http://www.aiai.org/ocs/index.php/ICWSM/09/paper/view/154>
- [38] M. Fazli, R. Bertram, and D. A. Striegel, "Multi-layer bundling as a new approach for determining multi-scale correlations within a high-dimensional dataset," *Bull. Math. Biol.*, vol. 86, no. 9, p. 105, Sep. 2024.
- [39] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. (1992). *Breast Cancer Wisconsin (Diagnostic) Data Set*. UCI Machine Learning Repository. [Online]. Available: <http://archive.ics.uci.edu/ml/>
- [40] E. Alpaydin and C. Kaynak, "Optical recognition of handwritten digits data set," UCI Mach. Learn. Repository, 1998, doi: [10.24432/C50P49](https://doi.org/10.24432/C50P49).
- [41] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021.
- [42] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proc. IEEE 7th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2020, pp. 747–748.
- [43] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist. Simul. Comput.*, vol. 3, no. 1, pp. 1–27, 1974.
- [44] J. M. Santos and M. J. Embrechts, "On the use of the adjusted Rand index as a metric for evaluating supervised classification," in *Proc. Int. Conf. Artif. neural Netw.*, 2009, pp. 175–184.
- [45] E.-R. Ardelean, R. L. Portase, R. Potolea, and M. Dinşoreanu, "A path-based distance computation for non-convexity with applications in clustering," *Knowl. Inf. Syst.*, vol. 67, no. 2, pp. 1415–1453, Feb. 2025.
- [46] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, Sep. 2018, doi: [10.21105/joss.00861](https://doi.org/10.21105/joss.00861).
- [47] F. Baffour-Awuah Jr., "Identifying distinct patient subgroups in sepsis," Ph.D. dissertation, Dept. Math., Florida State Univ., Tallahassee, FL, USA, 2025.
- [48] R. Vijay, P. Mahajan, and R. Kandwal, "Hamming distance based clustering algorithm," *Int. J. Inf. Retr. Res.*, vol. 2, no. 1, pp. 11–20, Jan. 2012.
- [49] Ö. Akay and G. Yüksel, "Clustering the mixed panel dataset using Gower's distance and k-palgorithms," *Commun. Statist. Simul. Comput.*, vol. 47, no. 10, pp. 3031–3041, Nov. 2018.



include infectious disease modeling, mathematical and computational biology, data assimilation, and epidemiology.



MEHRAN FAZLI received the B.S. degree in cell and molecular biology and the M.S. degree in applied mathematics from the University of Tehran and the Ph.D. degree in mathematics from Florida State University, where his research centered on multi-time scale analysis and synchronization in endocrine cells. He is currently affiliated with The Henry M. Jackson Foundation for the Advancement of Military Medicine Inc., in support of the ACESO Initiative. He develops computational tools, such as multi-layer bundling (MLB) and dynamical systems cost-fitting landscape (DSCL), to analyze high-dimensional biological data. He is a Biomathematician, with a focus on mathematical modeling, dynamical systems theory, and machine learning to explore complex biological phenomena, focusing on sepsis and immune system dynamics.



DEBORAH A. STRIEGEL received the Ph.D. degree in mathematics from Florida State University. She is currently a Biomathematician, who works as the Biomath Research and Development Lead of the Austere environments Consortium for Enhanced Sepsis Outcomes (ACESO), focusing on developing mathematical tools to advance sepsis understanding. Prior to working at ACESO, she was a Postdoctoral Research Fellow at the Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), NIH, Bethesda, MD, USA.

...